Acta Baltico-Slavica

Acta Baltico-Slavica, 45
Warszawa 2021

Article No. 2576
DOI: 10.11649/abs.2576
Citation: Roszko, R. (2021). O nowych ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusach równoległych oraz ich zastosowaniach. *Acta Baltico-Slavica, 45*, Article 2576. https://doi.org/10.11649/abs.2576

Roman Roszko
Instytut Slawistyki Polskiej Akademii Nauk
Warszawa
roman.roszko@ispan.waw.pl
https://orcid.org/0000-0002-2291-6939

O nowych ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusach równoległych oraz ich zastosowaniach

1. Wstęp

Dwudziesty pierwszy wiek przyniósł badaczom szeroko rozumianych nauk humanistycznych i społecznych wiele cyfrowych zasobów i narzędzi językowych, które w znaczącym stopniu przyczyniają się do skoku jakościowego i ilościowego prowadzonych obecnie badań. Dowodem rosnącego zainteresowania lingwistyką cyfrową są liczne publikacje. Przykład analiz leksykologicznych stanowią prace W. Sosnowskiego, J. Satoły-Staśkowiak, *A contrastive analysis of feminitives in Bulgarian, Polish and Russian* (Sosnowski & Satoła-Staśkowiak, 2019) oraz D. Blagoewej, M. P. Jaskota, W. Sosnowskiego *A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology* (Blagoeva i in., 2019). Z obszaru wielojęzyczności można wskazać pracę J. Fellerera *Urban multilingualism in East-Central Europe: The Polish dialect of Late-Habsburg Lviv* (Fellerer, 2020). Z kolei A. Wawer w pracy *Sentiment analysis*

Publisher: Institute of Slavic Studies, Polish Academy of Sciences
[Wydawca: Instytut Slawistyki Polskiej Akademii Nauk]

for Polish podejmuje się badań nad sentymentem w języku polskim (Wawer, 2019). Przykładem badań nad wzorcami relacji leksykalno-semantycznych jest artykuł A. Dziob i M. Piaseckiego *Dynamic verbs in the Wordnet of Polish* (Dziob & Piasecki, 2018). Natomiast M. Maziarz i E. Rudnicka rozważają kwestię rozszerzenia WordNetów (słowosieci) o glosy i relacje wieloznaczne na potrzeby rozpoznawania zjawisk, będących skojarzeniami zmysłowymi wykraczającymi poza typowe leksykalne relacje semantyczne (Maziarz & Rudnicka, 2020). Z zakresu przekładu można wymienić prace Ł. Grabowskiego *A corpus-driven study of translational and non-translational texts: The Case of Nabokov's 'Lolita'* (Grabowski, 2012) oraz E. Kaczmarskiej *Metody ustalania ekwiwalentów czasowników wyrażających stany emocjonalne w przekładzie czesko-polskim na materiale z korpusu równoległego InterCorp* (Kaczmarska, 2019). Przykładem badań psychologicznych jest wieloautorska praca I. Kaźmierczak, J. Sarzyńskiej-Wawer, A. Wawra i M. Chądzyńskiej (Kaźmierczak i in., 2020), w której autorzy podejmują zagadnienie opisu krytycznych wydarzeń życiowych i ich psychologicznych konsekwencji na podstawie rodzaju języka używanego przez pacjentów cierpiących na depresję i jego związku z rozwojem osobowości.

W ostatnim czasie powstają nie tylko cyfrowe zasoby językowe i narzędzia do automatycznego przetwarzania języka, lecz także rozbudowane infrastruktury badawcze, które łączą rozproszone zasoby i narzędzia na jednej platformie (por. Piasecki i in., 2018). Łączenie zasobów i narzędzi wbrew pozorom nie jest procesem czysto mechanicznym. Towarzyszy temu nieustanne uaktualnianie i wyrównywanie wszystkich zasobów i narzędzi do najnowszych wspólnych standardów. W roku 2020 ruszyła realizacja wielkiego projektu CLARIN-PL-BIZ w ramach „Programu Operacyjnego Inteligentny Rozwój 2014–2020" (POIR 4.2) w osi priorytetowej IV „Zwiększenie potencjału naukowo-badawczego" i działaniu 4.2 „Rozwój nowoczesnej infrastruktury badawczej sektora nauki". Jego celem jest utworzenie platformy badawczo-rozwojowej do przetwarzania języka naturalnego i eksploracji dużych zasobów danych językowych zapewniających dostęp do technologii językowych oraz mechanizmów ich łączenia z myślą o konstrukcji systemów analizy tekstów dla języka polskiego oraz pozostałych języków europejskich oraz hebrajskiego. Należy tu wyjaśnić, że niewielkie zbiory danych językowych można przetwarzać z zastosowaniem powszechnie dostępnych metod, jakie mogą być oferowane nawet w znanych edytorach tekstów czy popularnych przeglądarkach wielojęzycznych. Natomiast duże zbiory danych językowych wymagają zupełnie innych narzędzi, które są w stanie analizować te dane i jednocześnie dostarczać nowej informacji na ich temat. Analizy dużych zbiorów danych eliminują błędy, które są typowe dla analiz małych zasobów, tzw. próbek.

W ramach powstającej platformy badawczo-rozwojowej przewidziano:

- utworzenie centrum technologicznego (CTech), stanowiącego bazę dla technologii eksploracji danych językowych;

- zastosowanie w CTech zaawansowanych technologii językowych do inteligentnego przetwarzania wielkich niejednorodnych danych na nowych płaszczyznach, niewspieranych przez powstałe dotychczas infrastruktury badawcze ani technologie;
- opracowanie i wdrożenie odpowiednich standardów konstrukcji zasobów i narzędzi językowych;
- opracowanie i wytworzenie nowych narzędzi analizy danych językowych, działających jako jedna spójna i prosta w swej istocie struktura;
- przygotowanie i dostarczenie nowych danych ręcznie znakowanych do badań i trenowania narzędzi językowych;
- unifikacja zasobów i narzędzi językowych dla języków europejskich do postaci interoperacyjnej;
- zapewnienie użytkownikom CTech pełnego i łatwego dostępu do archiwów zawierających zasoby, narzędzia i technologie językowe.

Konstrukcja CTech przewidziana jest do roku 2024. Poszczególne etapy zadań są związane z:

a) utworzeniem systemu do gromadzenia i przechowywania danych językowych;
b) dostosowaniem narzędzi językowych do standardów komercyjnych i rozszerzeniem ich funkcjonalności;
c) utworzeniem kluczowych zasobów językowych dla języka polskiego połączonych z zasobami angielskimi, bułgarskimi, litewskimi, słoweńskimi, rosyjskimi i in. oraz z Linked Open Data[1];
d) opracowaniem narzędzi do analizy wydźwięku oraz emocji;
e) konstrukcją środowiska informatycznego do tworzenia systemów dialogowych;
f) wytworzeniem narzędzi do wydobywania informacji z danych tekstowych, w tym bazujących na semantycznej analizie tekstu i elementach analizy dyskursu;
g) opracowaniem ogólnego systemu do odpowiadania na pytania w języku naturalnym.

W punkcie c) opisane zadanie związane z utworzeniem kluczowych zasobów językowych dla języka polskiego połączonych z zasobami bułgarskimi, litewskimi, słoweńskimi i rosyjskimi obejmuje między innymi zaprojektowanie i konstrukcję

[1] Ten termin łączy w sobie dwa pojęcia: *danych otwartych* (Open Data) i *danych połączonych* (Linked Data). O ile pojęcie *danych otwartych* jest oczywiste, o tyle *danych połączonych* wymaga krótkiego wyjaśnienia. Za dane połączone uważa się kolekcje różnych zbiorów danych powiązanych ze sobą strukturalnie w jedną sieć.

czterech ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych (polsko-bułgarskiego, polsko-litewskiego, polsko-słoweńskiego i polsko-rosyjskiego).

2. Cel konstrukcji ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych

Głównym celem utworzenia czterech wyżej wymienionych zbiorów danych jest udostępnienie przedstawicielom nauk humanistycznych i społecznych oraz programistom zajmującym się projektowaniem i wytwarzaniem narzędzi do przetwarzania języka naturalnego (NLP)[2] wysokiej jakości ręcznie zrównoleglonych i znakowanych równoległych dwujęzycznych zasobów polsko-bułgarskich, polsko-litewskich, polsko-rosyjskich i polsko-słoweńskich.

Ci pierwsi odbiorcy uzyskają dostęp do tych korpusów w rozbudowanej przyjaznej użytkownikowi wielojęzycznej przeglądarce webowej *KonText*[3] (*KonText*, b.d.; Machálek, 2020). Dzięki skrupulatnemu doborowi utworów reprezentujących możliwie najszerszy zestaw stylów językowych, zaawansowanemu wstępnemu przetworzeniu i oczyszczeniu tekstów[4] oraz ręcznie przeprowadzonym segmentacji i wielo-

[2] Przykłady narzędzi NLP: systemy do przechowywania i udostępniania danych językowych, wyszukiwarki korpusowe, analizatory cech gramatycznych, składniowych, stylometrycznych, anotatory znakujące/kodujące zasoby językowe (np. tagery, lematyzatory), syntezatory mowy, systemy do przetwarzania mowy itd.

[3] *KonText* jest to zaawansowana nieustannie modernizowana webowa wyszukiwarka korpusowa udostępniająca użytkownikowi szereg sposobów przeszukiwania zasobów, w tym zaawansowany z zastosowaniem języka zapytań CQL (ang. Corpus Query Language) obsługującego atrybuty, operatory, wyrażenia regularne, klasy i fleksemy słów, kategorie gramatyczne i metaanotację. *KonText* zapewnia szczegółowe profilowanie (sortowanie, filtrowanie) wyszukiwanych konkordancji zarówno w lewym jak i prawym kontekście od wyszukanej formy (KWIC: Key Word In Context). Ponadto umożliwia obliczenie szeregu automatycznie generowanych miar i zastosowanie wielu innych przydatnych opcji i narzędzi. Na stronach Clarin-PL jest dostępna instrukcja do wyszukiwarki korpusowej *KonText* https://nextcloud.clarin-pl.eu/index.php/s/fzAZg9xbxA4YEdu oraz ściągawka dla instrukcji do *KonTextu* https://nextcloud.clarin-pl.eu/index.php/s/IsIriR9v5Hopaml#pdfviewer.

[4] Zaawansowane wstępne przetworzenie tekstów obejmuje korektę pisowni (w tym usunięcie ligatur, znaków technicznych, korektę skrótów, dodanie wymaganych lub usunięcie zbędnych spacji itd.), wyróżnienie najmniejszych jednostek czyli tokenów. W przypadku części filmowych list dialogowych wstępne przetworzenie tekstów obejmuje korektę tłumaczenia i łagodne uzgodnienie treści obu wersji językowych, np. PL *Byliśmy na wycieczce w Universal Studio, a potem w woskowym muzeum Hollywood.* || UK *Ми ходили на екскурсію по кіностудії Довженка, а потім Чарлі водив мене у музей воскових фігур Верховної Ради.* → PL *Byliśmy na wycieczce w Universal <u>Studios</u>, a potem w <u>Muzeum Figur Woskowych</u> w Hollywood.* || UK *Ми ходили на екскурсію по <u>Студії Universal</u>, а потім Чарлі водив мене в <u>Голлівудський музей воскових фігур</u>.*

poziomowej anotacji te korpusy staną się wartościowym obiektem analiz, w których znajdą zastosowanie następujące metody badawcze (por. Kaczmarska, 2019):

i. sterowane korpusem (ang. corpus-driven approach);
ii. oparte na korpusie (ang. corpus-based lub corpus-supported approach);
iii. ilustrowane przykładami korpusowymi (corpus-illustrated lub corpus-informed approach);
iv. mieszane, łączące analizę danych korpusowych i niekorpusowych (takich jak wywiady, ankiety) (ang. corpus-assisted analysis);
v. generowane korpusem (ang. corpus-induced approach).

Kolejność wyszczególnienia metod badawczych odzwierciedla typowe zastosowania ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych w naukach humanistycznych i społecznych. W pierwszej kolejności te korpusy są przeznaczone do badań teoretycznych (por. Koseska i in., 2007, 2009) i statystycznych (ilościowe połączone z analizą jakościową, np. częstość użycia zaimka osobowego w funkcji wykładnika określoności w językach polskim i litewskim, por. Roszko, 2015), w dalszej zaś – do typowych badań materiałowych i komercyjnych badań stosowanych. Wymieniona w punkcie *i* metoda opisuje badania prowadzące do formułowania nowych teorii. W tej metodzie odrzuca się wstępne formułowanie hipotez (por. Grabowski, 2012, 2015, ss. 28–29). W punkcie *ii* mowa jest o badaniach, w których wyekscerpowane z korpusów fakty służą weryfikacji uprzednio znanej lub na cele prowadzonych badań postawionej hipotezy (por. Grabowski, 2015, ss. 28–29). Badania ilustrowane przykładami korpusowymi (punkt *iii*) są kolejnym szczególnym zastosowaniem ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych[5], które sprawdzą się jako przybliżony[6] tzw. zbiorowy wyważony informator cyfrowy. Wartość argumentacyjna dowodów bazujących na ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusach równoległych jest siłą rzeczy wyższa od tych, których podstawą są częściowo lub całkowicie automatycznie generowane korpusy dwujęzyczne. Rozszerzeniem przedstawionej w punkcie *iii* metody jest wariant badań ilustrowanych przykładami korpusowymi (punkt *iv*), w którym analizy korpusowe są tylko jednym z wielu źródeł danych (por. Hebal-Jezierska, 2013).

[5] Dane oraz miary automatycznie obliczone na bazie ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych nie wymagają dodatkowej weryfikacji materiału będącego podstawą tych obliczeń. W korpusach automatycznie zrównoleglonych i znakowanych zawsze zachodzi obawa uwzględnienia w wynikach błędnie zidentyfikowanych form językowych. Ponadto należy mieć na uwadze, że w niektórych korpusach nie przeprowadza się weryfikacji pisowni, a nawet języka. Na przykład w korpusie *OPUS* dochodzi do mylenia języków zapisanych cyrylicą (np. rosyjskojęzyczne teksty są uznane za ukraińskie), a w korpusie *InterCorp* niektóre bułgarskie utwory zostały włączone do zasobów rosyjskich.

[6] Określenie „przybliżony" jest konieczne, bowiem korpus konstruowany w części na tłumaczeniach, a nie tylko na tekstach naturalnie powstałych w danym języku, jest pewnym przybliżeniem faktycznego języka. Tłumaczenia mogą zawierać elementy (leksykalne, składniowe itd.) generowane językiem oryginału.

W analizach i zastosowaniach generowanych korpusem (punkt *v*; por. Reynaert, 2006) zastosowanie ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych jest bodaj najmniejsze, bowiem do tego typu analiz wymagane są bardzo duże korpusy (*big data*). Dobrym przykładem takiego korpusu jest rozwijany od 2010 roku jednojęzyczny korpus monitorujący *MoncoPL* (Pęzik, 2020, http://monco.frazeo.pl/)[7]. Takie właśnie korpusy pozwalają na automatyczną analizę jakościowo-ilościową z myślą o konkretnych zastosowaniach praktycznych, także komercyjnych. Należy mieć na uwadze, że badania generowane korpusami wielojęzycznymi, najczęściej noszą charakter stosowany, np. na rzecz rozwoju przekładu maszynowego.

Drugi rodzaj odbiorców – programiści zajmujący się projektowaniem i wytwarzaniem narzędzi NLP – widzą w ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusach równoległych źródło do trenowania i testowania nowych oraz doskonalenia istniejących narzędzi językowych. Każdy tego typu korpus dostarcza trzech rodzajów danych: dla języka A, dla języka B oraz dla obu wzajemnie powiązanych ze sobą języków. Warto podkreślić, że polskojęzyczne zasoby wszystkich czterech przygotowywanych korpusów mogą być dowolnie łączone, co wydatnie zwiększy objętość ręcznie znakowanych zasobów dla języka polskiego.

Ręcznie zrównoleglone i znakowane korpusy mogą zostać zastosowane w wypracowaniu algorytmów projekcji/rzutowania znaczeń z jednego języka na drugi. Szczególnie chodzi tu o projekcję tych znaczeń, które w jednym języku są jednoznacznie wyrażane na płaszczyźnie formalnej, w drugim zaś – dochodzi do tzw. niedopowiedzenia językowego (Koseska & Roszko, 2015), por. użycia polskiego *kiedyś* w dwóch różnych znaczeniach kwantyfikacyjnych *kiedyś przyszedł* i *kiedyś przyjdzie* wobec dwu litewskich form *kažkada* i *kada nors* (litew. *kažkada atėjo* [kiedyś przyszedł] a *kada nors ateis* [kiedyś przyjdzie]). Można również zakładać, że ręcznie zrównoleglone i znakowane korpusy równoległe mogą być pomocne w doskonaleniu narzędzi do automatycznego zrównoleglania zasobów dwujęzycznych.

Ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe są idealnymi zasobami na rzecz rozwoju rekurencyjnych sieci neuronowych, w którego efekcie dochodzi do doskonalenia przekładu maszynowego i rozwoju sztucznej inteligencji. Jednak należy podkreślić, że te korpusy nie mogą być wyłącznym źródłem zasobów treningowych algorytmów neuronowego tłumaczenia maszynowego ze względu na ograniczoną objętość.

[7] Korpus monitorujący to nieustannie rozwijany otwarty zasób jednojęzyczny. Jego główne cechy to diachroniczny charakter, referencyjność w ograniczeniu do próbkowanych rejestrów oraz duże tempo przyrostu (Leech, 2002; Pęzik, 2020). Przedstawiony tu jako przykład korpus *MoncoPL* dziennie zwiększa swą objętość o blisko 1,65 mln słów. Jest on uzupełniany, tagowany i lematyzowany automatycznie. Żadne dwujęzyczne korpusy równoległe, nawet te potencjalne utworzone ze wszystkich dostępnych tekstów równoległych dla obu języków nie mogą konkurować objętością z jednojęzycznymi korpusami monitorującymi.

3. Ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe a inne równoległe korpusy dwu- i wielojęzyczne

Podstawowe różnice między znanymi korpusami wielojęzycznymi, takimi jak *InterCorp* (Čermák & Rosen, 2012), *Parallel Corpora of the Russian National Corpus* (Добровольский i in., 2005), *ParaSol* (von Waldenfels, 2011), *Polish–Russian Parallel Corpus* (Łaziński & Kuratczyk, 2016), *Opus* (Tiedemann, 2016) itd. a powstającymi w ramach projektu CLARIN-PL (CLARIN-PL, b.d.) ręcznie zrównoleglonymi i znakowanymi dwujęzycznymi korpusami równoległymi to przede wszystkim charakterystyczne dla tych ostatnich: staranny dobór zasobów z oceną merytoryczną tekstu i jakości samego języka, wstępne zaawansowane przetwarzanie zasobów, ręczne zrównoleglenie i ręczne znakowanie, ponadto wprowadzenie rozbudowanych metadanych i zastosowanie najnowszych standardów opisu zasobów i zapisu danych. Na każdym kluczowym etapie konstrukcji tych korpusów przeprowadzana jest kontrola jakościowa realizowanych zadań. Na przykład, zasoby każdego języka są niezależnie opisywane przez dwóch anotatorów. Wyniki ich pracy są przedstawiane trzeciemu (super)anotatorowi, który decyduje o ostatecznym wyniku ręcznego znakowania. Ręczne zrównoleglenie i znakowanie są czasochłonne i kosztowne. Dlatego konstruowane cztery ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe będą miały zdecydowanie mniejszą objętość niż większość wyżej w tym akapicie wymienionych korpusów równoległych. Przyjęto, że objętość każdego korpusu osiągnie 1 mln słowoform dla każdego języka. Ukończenie prac nad tymi korpusami jest planowane na rok 2024.

Rzeczone korpusy będą udostępnione odbiorcom w dwóch wariantach. W wersji do pobrania składającej się z szeregu plików zapisanych w aktualnych w momencie publikacji tych korpusów standardach. Ponadto z myślą o użytkownikach – badaczach z szeroko rozumianych nauk humanistycznych i społecznych – zasoby będą dostępne w przeglądarce webowej *KonText*. Jak wiadomo, niektóre korpusy równoległe nie posiadają interfejsu użytkownika, np. *PELCRA Polish–Russian parallel corpus* (*PELCRA*, b.d.). Warto również wspomnieć, że konstrukcja części korpusów wielojęzycznych przebiega automatycznie bez udziału i nadzoru człowieka. Wyjątkiem może tu być *InterCorp* (Čermák & Rosen, 2012), w którym segmentacja części zasobów wielojęzycznych została przeprowadzona ręcznie i oznaczona jako *Core* oraz korpusy wielojęzyczne CLARIN-PL (Duszkin i in., 2021), w których segmentacja zasobów przebiegała w dwóch etapach: po początkowym maszynowym zrównolegleniu została przeprowadzona ręczna korekta. Możliwym niezamierzonym rezultatem automatycznej segmentacji zasobów są różnej wagi błędy we wzajemnym przyporządkowaniu segmentów w poszczególnych parach językowych. Ponadto niektóre korpusy wielojęzyczne nie są znakowane. Oznacza to istotne zubożenie funkcjonalności wyszukiwania interesujących użytkownika form, wyrażeń. Znakowanie zasobów

korpusowych obejmuje między innymi tagowanie warstw morfologicznej, składniowej i semantycznej oraz lematyzację. We wszystkich znanych mi niekomercyjnych korpusach wielojęzycznych tagowanie i lematyzacja zostały przeprowadzone automatycznie. Ten fakt pociąga za sobą znaczny odsetek źle zinterpretowanych form. Poniżej przedstawiam wybrane przykłady błędnego znakowania w *Polsko-Rosyjskim Korpusie Uniwersytetu Warszawskiego* (Łaziński & Kuratczyk, 2016)[8]. W ścieżce wyborów przy konstrukcji zapytania dla zasobów polskojęzycznych po wybraniu: [Wyszukiwanie morfologiczne] – [Część mowy] – wybór "Rzeczownik" – [Kategorie gramatyczne] – wybór "Wołacz" uzyskałem odpowiedź (http://www.pol-ros.polon.uw.edu.pl/searchresults/searchmpl.php?stringpl=&subst=on&depr=on&voc=on&limitPl=10)[9] zawierającą między innymi takie zdania:

1) *Podporucznik **Tadzio** Jarzębski, piastujący stanowisko podkomisarza policji, punktualnie przybył na umówione miejsce.*
2) *To waży sto **kilo**!*
3) *Drzwi otworzyły się przed nim, zanim zdążył przyłożyć **palecdo** dzwonka* [sic!].
4) *Noszę nazwisko po jej **synu**, a na imię mam tak samo jak ona.*

Na początku należy zaznaczyć, że w odpowiedzi nie otrzymuje użytkownik żadnego graficznego wyróżnienia wyszukanych form. Wytłuszczeniem zaznaczyłem te formy, które moim zdaniem mogły zostać błędnie zinterpretowane w automatycznym procesie tagowania. W zdaniu pierwszym domniemaną formą wołaczową jest *Tadzio*. Domyślam się, że tej słowoformie został błędnie przypisany potencjalny lemat *Tadzia*, a następnie, zgodnie z paradygmatem jednak twardotematowym, fleksja wołacza *-o*. W drugim przykładzie zakładam, że formą wołaczową jest nieodmienny rzeczownik *kilo*. Możliwe, że nieodmienność tego leksemu oraz następujący po nim wykrzyknik mógł spowodować, że w automatycznym tagowaniu tej formie przypisano wartość wołacza. W kolejnym trzecim zdaniu trudno wskazać formę wołacza. Możliwe, że tager ustalił formę *palecdo* [sic!] jako wołaczową. W czwartym zdaniu formą zinterpretowaną jako wołacz jest zapewne *synu*. Jest to postać wspólna dla miejscownika i wołacza. Można przypuszczać, że algorytm interpretacji formy *synu* jako wołaczowej jest wynikiem koincydentacji samej postaci *synu* jako potencjalnie wołaczowej oraz następującego po niej przecinka, który zgodnie z normami językami polskiego stosuje się po zwrotach w wołaczu.

[8] Wybór tego korpusu jest zupełnie przypadkowy. Analogiczne odpowiedzi do przedstawianych tu przykładów błędnej interpretacji form są właściwe wszystkim cytowanym w tym artykule korpusom.

[9] W pierwszej dziesiątce odpowiedzi tylko w czterech zdaniach stwierdziłem użycie wołacza, dwukrotnie *Chryste Panie* oraz pojedynczo *Matko Boska* i *Panie* (voc. sg. masc.). Łączna liczba segmentów, w których ma być notowany wołacz, wynosi 23.764, co w odniesieniu do objętości korpusu (łącznie 30 mln polskich i rosyjskich słowoform) wydaje się liczbą nieprawdopodobną.

Przedstawione tu wybrane przykłady możliwych błędów w automatycznie lub półautomatycznie konstruowanych korpusach, nie negują użyteczności tak utworzonych korpusów. Istotną wskazówką dla potencjalnego użytkownika może być przede wszystkim czas opublikowania korpusu oraz data ostatniej aktualizacji. Czym starszy korpus i/lub dużo czasu upłynęło od jego aktualizacji, tym prawdopodobieństwo wystąpienia błędów jest wyższe. Jest to pochodną użycia do automatycznego znakowania zasobów starszych mniej doskonałych narzędzi. Niezależnie od potencjalnych błędów takie korpusy są użyteczne do prowadzenia wielu badań. Co innego, gdy taki korpus ma służyć trenowaniu narzędzi. Wówczas niezastąpione są ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe. Są one podstawą do konstrukcji wielu niezbędnych w nieliniowym przetwarzaniu języka naturalnego narzędzi.

4. Zasady konstrukcji ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych

W realizowanym projekcie CLARIN-PL w początkowym okresie (IV kwartał 2020 – początek 2021) opracowano metodologię konstrukcji ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych. Nim jednak doszło do sformułowania założeń konstrukcji tego typu korpusów, gruntownie zapoznano się z istniejącymi korpusami jednojęzycznymi i wielojęzycznymi równoległymi i porównawczymi, zwłaszcza z tymi, w których są reprezentowane języki bałtyckie i słowiańskie. Wyniki przeprowadzonej eksploracji „rynku korpusowego" zamieszczono w czterech obszernych zamkniętych raportach CLARIN-PL-BIZ autorstwa Jakuba Banasiaka (IS PAN), Pawła Kowalskiego (IS PAN), Danuty Roszko (UW) i Romana Roszko (IS PAN). W tych raportach również zawarto szczegółową informację o rynku komercyjnych i znajdujących się w wolnym dostępie narzędzi językowych, wskazano potencjalne źródła wielojęzycznych zasobów i mechanizmy automatycznego pozyskiwania takich danych, np. *LAMBERT* (Garncarek i in., 2021). Dokonano ewaluacji dostępnych tagerów dla języków słowiańskich i bałtyckich. Zapoznano się również z literaturą przedmiotu, w której krytycznie odnoszono się do już powstałych korpusów (Charciarek, 2018, 2019a, 2019b; Piotrowski & Grabowski, 2013), w tym też z pracami zbiorowymi poruszającymi temat korpusów wielojęzycznych (Gruszczyńska & Leńko-Szymańska, 2016; Hebal-Jezierska, 2014 i wiele innych). Sporo uwagi poświęcono rynkowi użytkowników takich korpusów. Tu głównie wykorzystano dane Centrum Wiedzy PolLinguaTec CLARIN-PL, do którego zgłaszają się użytkownicy (również potencjalni) i przedstawiają swoje oczekiwania wobec infrastruktury badawczej. Nie mniejszą rolę przywiązano do sugestii i wskazań Zespołu Oceniającego Ministerstwa Edukacji i Nauki (powołanego przez MEiN) oraz prężnie działającej Rady Programowej CLARIN-PL, powołanej przez konsorcjum

CLARIN-PL na polecenie MEiN. Wyznaczone w wyżej opisanym procesie potencjalne prace projektowe zostały ostatecznie sprecyzowane w bezpośrednim kontakcie z przedstawicielami innych centrów technologii językowych (w Bułgarii, Czechach, Litwie, Rosji i Słowenii[10]) mniej lub bardziej związanych z CLARIN-ERIH. Ostatecznie dokonano wyboru języków, które na pierwszym etapie prac zostaną włączone do przygotowywanych ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych. Ustalono poniższe pary językowe: polsko-bułgarską, polsko-litewską, polsko-rosyjską i polsko-słoweńską. Dokonany przez zespół CLARIN-PL wybór dobrze koresponduje z językoznawczym stanowiskiem o różnorodności typologicznej tych języków. Zauważmy, z jednej strony włączony został syntetyczny język litewski o stosunkowo zachowawczej i przejrzystej budowie morfologicznej, dużej regularności gramatykalizacji znaczeń, z drugiej zaś – analityczny i innowacyjny (rodzajnik, struktury aspektualno-temporalne, utrata deklinacji przy zachowaniu form wołacza i inne wspólne dla bałkańskiej ligi językowej cechy) język bułgarski. Typologicznie języki rosyjski i polski są sytuowane pomiędzy syntetycznym językiem litewskim i analitycznym językiem bułgarskim.

Na etapie wstępnym prac nad tymi korpusami ustalono zasady selekcji utworów, ustalono wzorzec wewnętrznego zrównoważenia dla poszczególnych korpusów.

Poniżej przedstawiam kluczowe etapy prac, których wynikiem będą pierwsze cztery ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe języków słowiańskich i bałtyckich:

1. selekcja próbek (tekstów) z dbałością o poprawność językową, właściwe wewnętrzne zrównoważenie zasobów i zbalansowaną reprezentację rejestrów;
2. nadzorowane automatyczne wstępne przetwarzanie tekstów (obejmuje między innymi czyszczenie, korektę błędów pisowni i interpunkcji oraz wstępną tokenizację);
3. ręczna segmentacja na poziomie zdania w parach językowych w oparciu o ściśle zdefiniowane wzorce wyróżniania i uzgadniania międzyjęzykowych odpowiedniości;
4. ręczna segmentacja na poziomie najmniejszych wyróżnianych jednostek, tzw. tokenów, prowadzona dla każdego języka niezależnie;
5. ręczna dwuetapowa lematyzacja tokenów, polegająca na niezależnym opisie form każdego języka przez dwóch anotatorów (etap 1), a następnie na zestawieniu tych wyników i, w przypadku rozbieżnego znakowania, wskazaniu wersji ostatecznej przez superanotatora (etap 2);
6. ręczne dwuetapowe znakowanie warstwy fleksyjnej (por. wyżej p. 5);

[10] Współpraca z ośrodkami zagranicznymi opiera się na obustronnej wymianie lub jednostronnym wsparciu w zakresie istniejących i powstających technologii językowych oferowanych przez te ośrodki, przydatnych do realizacji opisywanych tu zadań projektowych CLARIN-PL-BIZ.

7. przygotowanie rozbudowanego opisu zasobów (metadane);
8. konwersja danych do obowiązujących standardów (dla zastosowań treningowych) oraz przygotowanie wersji zgodnej z *KonTextem*;
9. rozbudowa interfejsu *KonTextu*, obejmująca między innymi utworzenie konstruktora zapytań oraz rozbudowanej sekcji wyboru zasobów na podstawie wartości zdefiniowanych w metadanych.

Rozważane jest pozyskiwanie quasi-równoległych danych w automatycznym przeszukiwaniu zasobów sieciowych z wykorzystaniem narzędzi LASER czy LAMBERT. Decyzja o włączeniu tego typu danych jeszcze nie zapadła. Za jej włączeniem przemawia możliwość pozyskania do korpusów doniesień prasowych, które w całości lub w części są tłumaczone na wiele języków. Wśród doniesień agencyjnych zdarzają się również tłumaczenia wypowiedzi konkretnych osób (polityków, przywódców duchowych itd.). Za niewłączaniem tego typu tekstów przemawia konieczność wydzielenia tych zasobów w osobnej części korpusu, quasi-równoległy charakter tego typu wiadomości oraz duża liczba krótkich/hasłowych tekstów (cytatów). Zastosowanie takich danych z natury nieciągłych w udoskonalaniu maszynowego tłumaczenia jest mniej skuteczne niż w przypadku użycia właściwych zasobów równoległych.

5. O wpływie ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusów równoległych na badania z zakresu szeroko rozumianych nauk humanistycznych i społecznych

Po zakończeniu prac nad ręcznie zrównoleglonymi i znakowanymi dwujęzycznymi korpusami równoległymi ich zasoby zostaną udostępnione między innymi w webowej przeglądarce *KonText*. To ten wariant publikacji korpusów jest przede wszystkim kierowany do użytkowników reprezentujących nauki humanistyczne i społeczne. Opisana w przypisie 3. funkcjonalność przeglądarki *KonText* w połączeniu z przygotowanymi z dbałością o możliwie najwyższą jakość zasobów umożliwi językoznawcom prowadzenie wyczerpujących badań i analiz.

Warto też wspomnieć o pośrednim zastosowaniu tych korpusów w badaniach humanistycznych i społecznych. Mam tu na myśli wszelkie nowe zasoby oraz narzędzia i programy przetwarzania języka naturalnego, do których powstania po części przyczynią się te właśnie ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe. Jak już wspominałem, nadrzędnym celem konstrukcji tych korpusów jest doskonalenie modeli międzyjęzykowych, które wpłyną na poprawę istniejących oraz utworzenie nowych narzędzi NLP. Spodziewać się można, że wytrenowane i/lub testowane na tych korpusach narzędzia będą szeroko stosowane przez badaczy nauk humanistycznych i społecznych. Nie można też zapominać o decydującym wpływie

modeli międzyjęzykowych na poprawę przekładu maszynowego oraz doskonalenia zyskujących na popularności wielojęzycznych chatbotów do automatycznej obsługi klientów.

6. Podsumowanie

Opisane w tym artykule ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe powstają z myślą o dostarczeniu badaczom i programistom wzorcowych zasobów do badań naukowych (z zakresu nauk humanistycznych i społecznych) i prac projektowych w obszarze szeroko rozumianego przetwarzania języka naturalnego. W ramach przedstawionego projektu CLARIN-PL-BIZ są konstruowane cztery takie korpusy: polsko-bułgarski, polsko-litewski, polsko-rosyjski i polsko-słoweński. Cechą wyróżniającą te korpusy na tle innych dwujęzycznych korpusów równoległych jest odstąpienie od użycia jakichkolwiek narzędzi automatyzujących pracę. Wszystkie etapy konstrukcji tych korpusów są i będą realizowane przez zespół specjalistów-językoznawców. Zakładana objętość każdego korpusu wyniesie około miliona słowoform. Ręcznie zrównoleglone i znakowane zasoby polskie wszystkich czterech korpusów zostaną połączone i udostępnione jako kolejny ręcznie znakowany korpus języka polskiego. Obok znanych i szeroko stosowanych ręcznie znakowanego milionowego podkorpusu *NKJP* (Przepiórkowski i in., 2012) i nieustannie rozwijanego *Korpusu Języka Polskiego Politechniki Wrocławskiej* (Marcińczuk i in., 2015) będzie to trzeci potencjalny zasób treningowy. Ukończenie prac nad planowanymi czterema korpusami jest przewidywane w roku 2024. Zostaną one udostępnione w przeglądarce *KonText* na stronie CLARIN-PL oraz w postaci opisanych plików źródłowych w repozytorium *dSpace* także na stronie CLARIN-PL. Zasoby umieszczone w przeglądarce *KonText* są kierowane głównie do przedstawicieli nauk humanistycznych i społecznych. Natomiast odbiorcami plików źródłowych tych korpusów będą zespoły lingwistyczno-informatyczne CLARIN-PL pracujące nad konstrukcją nowych zaplanowanych i już opracowywanych narzędzi językowych dla języków polskiego, bułgarskiego, litewskiego, słoweńskiego i rosyjskiego. Otwarty charakter tych korpusów sprawi, że również inni twórcy oprogramowania będą mogli stosować te zasoby do trenowania i testowania własnych modeli międzyjęzykowych na rzecz nieustającego rozwoju sztucznej inteligencji.

Usługi oferowane przez CLARIN-PL są kompleksowe. Każdy zarejestrowany użytkownik może samodzielnie utworzyć własny korpus (korpusy) badawczy, deponując jego zasoby w bezpiecznej chmurze (*Clarin Cloud*, https://nextcloud.clarin-pl.eu/) i/lub w *Repozytorium* (*dSpace*, https://clarin-pl.eu/dspace/). Każdym zasobom użytkownik przypisuje stosowne prawa i decyduje o ewentualnym udostępnieniu innym badaczom (zdefiniowanym, konkretnym, grupom, wszystkim itd.). Wszystkie

zasoby umieszczone w *dSpace* gwarantują synchronizację z narzędziami oferowanymi przez CLARIN-PL oraz zewnętrznymi udostępnionymi w ramach europejskiej platformy CLARIN ERIH. Od lat działające Centrum Wiedzy PolLinguaTec służy natychmiastową pomocą użytkownikom. Na stronach CLARIN-PL przystępnie opisane są wszelkie oferowane zasoby (https://clarin-pl.eu/index.php/zasoby/) i narzędzia/usługi (https://clarin-pl.eu/index.php/uslugi/). W Mediatece (https://clarin-pl.eu/index.php/mediateka/) zamieszczono praktyczne instrukcje narzędzi, materiały warsztatowe, publikacje i prezentacje. Znajdują się tam również linki do wielu projektów z zakresu e-humanistyki w Polsce. Na stronach CLARIN-PL (https://clarin-pl.eu/dspace/) udostępniono użytkownikom wyszukiwarkę zasobów i narzędzi językowych. Warto odnotować istnienie kanału CLARIN-PL na YouTube (https://www.youtube.com/channel/UCqrhEITxu8_MIWPnFdYomPw/videos), gdzie dostępne są filmy instruktażowe oraz nagrania z konferencji i warsztatów.

Bibliografia

Blagoeva, D., Jaskot, M. P., & Sosnowski, W. (2019). A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology. *Cognitive Studies | Études cognitives*, *2019*(19), Article 1923. https://doi.org/10.11649/cs.1923

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427. https://doi.org/10.1075/ijcl.17.3.05cer

Charciarek, A. (2018). Možnosti využiti korpusu InterCorp v česko-polske překladové lexikografii. *Časopis pro moderní filologii*, *100*(2), 206–222.

Charciarek, A. (2019a). Korpus równoległy InterCorp w leksykografii przekładowej – możliwości i ograniczenia. *Roczniki Humanistyczne*, *67*(7), 79–92. https://doi.org/10.18290/rh.2019.67.7-5

Charciarek, A. (2019b). Využití paralelního korpusu v translatologii (na základě česko-polského InterCorpu). *Bohemistyka*, *2019*(2), 194–216. https://doi.org/10.14746/bo.2019.2.5

CLARIN-PL. (b.d.). *Polska infrastruktura CLARIN*. http://clarin-pl.eu/

Duszkin, M., Roszko, D., & Roszko, R. (2021). New parallel corpora of Baltic and Slavic languages – Assumptions of corpus construction. W K. Ekštein, F. Pártl, & M. Konopík (Red.), *Lecture Notes in Artificial Intelligence LNAI 12848: TSD 2021* (ss. 173–183). Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-83527-9_15

Dziob, A., & Piasecki, M. (2018). Dynamic verbs in the Wordnet of Polish. *Cognitive Studies | Études cognitives*, *2018*(18), Article 1718. https://doi.org/10.11649/cs.1718

Fellerer, J. (2020). *Urban multilingualism in East-Central Europe: The Polish dialect of late-Habsburg Lviv*. Rowman & Littlefield.

Garncarek, Ł., Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Turski, M., & Graliński, F. (2021). LAMBERT: Layout-aware language modeling for information extraction.

W J. Lladós, D. Lopresti, & S. Uchida (Red.), *Document Analysis and Recognition – ICDAR 2021* (ss. 532–547). Springer International Publishing. https://doi.org/10.1007/978-3-030-86549-8_34

Grabowski, Ł. (2012). *A corpus-driven study of translational and non-translational texts: The case of Nabokov's* Lolita. Wydawnictwo Uniwersytetu Opolskiego.

Grabowski, Ł. (2015). O frazeologii z perspektywy językoznawstwa korpusowego: Przegląd głównych nurtów badawczych ostatniego dwudziestolecia w Wielkiej Brytanii i USA. *Problemy Frazeologii Europejskiej*, *10*, 23–48.

Gruszczyńska, E., & Leńko-Szymańska, A. (Red.). (2016). *Polskojęzyczne korpusy równoległe*. Instytut Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Hebal-Jezierska, M. (2013). Podstawowe zasady korzystania z korpusów przy badaniu języka. W W. Chlebda (Red.), *Tropem korpusów: W poszukiwaniu optymalnych zbiorów tekstów* (ss. 17–30). Uniwersytet Opolski.

Hebal-Jezierska, M. (Red.). (2014). *Praktyczny przewodnik po korpusach języków słowiańskich*. Wydział Polonistyki Uniwersytetu Warszawskiego.

Kaczmarska, E. (2019). *Metody ustalania ekwiwalentów czasowników wyrażających stany emocjonalne w przekładzie czesko-polskim na materiale z korpusu równoległego InterCorp*. Wydział Polonistyki Uniwersytetu Warszawskiego.

Kaźmierczak, I., Sarzyńska-Wawer, J., Wawer, A., & Chądzyńska, M. (2020). Describing a critical life event and its psychological consequences: The type of language used by patients suffering from depression and its relationship with personality development. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*. https://doi.org/10.1007/s12144-020-00944-5

KonText – Corpus Query Interface. (b.d.). https://kontext.clarin-pl.eu/run.cgi/first_form

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Wydawnictwo Akademickie „Dialog".

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2009). Contrastive studies and semantic interlanguage. *Cognitive Studies | Études cognitives*, *2009*(9), 15–34.

Koseska, V., & Roszko, R. (2015). On semantic annotation in CLARIN-PL parallel corpora. *Cognitive Studies | Études cognitives*, *2015*(15), 211–236. https://doi.org/10.11649/cs.2015.016

Leech, G. (2002). The importance of reference corpora. W *Corpus lingüísticos: Presente y futuro* (ss. 1–11). Unibertsitate Zerbitzuetarako Euskal Ikastetxea.

Łaziński, M., & Kuratczyk, M. (2016). Korpus Polsko-Rosyjski Uniwersytetu Warszawskiego. W E. Gruszczyńska & A. Leńko-Szymańska (Red.), *Polskojęzyczne korpusy równoległe* (ss. 83–95). Instytut Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Machálek, T. (2020). KonText: Advanced and flexible corpus query interface. W *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (ss. 7003–7008). European Language Resources Association.

Marcińczuk, M., Oleksy, M., Kocoń, J., Bernaś, T., & Wolski, M. (2015). Towards an event annotated corpus of Polish. *Cognitive Studies | Études cognitives*, *2015*(15), 253–267. https://doi.org/10.11649/cs.2015.018

Maziarz, M., & Rudnicka, E. (2020). Expanding WordNet with gloss and polysemy links for evocation strength recognition. *Cognitive Studies | Études cognitives*, *2020*(20), Article 2325. https://doi.org/10.11649/cs.2325

PELCRA Polish-Russian parallel corpus. (b.d.). http://pelcra.pl/new/polrus

Pęzik, P. (2020). Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*, *7*, 133–150. http://doi.org/10.31261/FL.2020.07.11

Piasecki, M., Walkowiak, T., Rudnicka, E., & Bond, F. (2018). Lexical platform – the first step towards user-centred integration of lexical resources. *Cognitive Studies | Études cognitives*, *2018*(18), Article 1811. https://doi.org/10.11649/cs.1811

Piotrowski, T., & Grabowski, Ł. (2013). Interpretacja danych frekwencyjnych z korpusów językowych: Opis pewnych problemów (na kilku przykładach z życia wziętych). W W. Chlebda (Red.), *Na tropach korpusów: W poszukiwaniu optymalnych zbiorów tekstów* (ss. 59–71). Wydawnictwo Uniwersytetu Opolskiego.

Przepiórkowski, A., Bańko, M., Górski, R., & Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN.

Reynaert, M. (2006). Corpus-induced corpus clean-up. W N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Red.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC–2006, Trento* (ss. 87–92). European Language Resources Association (ELRA). https://aclanthology.org/L06-1000

Roszko, D. (2015). *Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej (na tle literackich języków polskiego i litewskiego)*. Instytut Slawistyki Polskiej Akademii Nauk. https://hdl.handle.net/20.500.12528/31

Sosnowski, W. P., & Satoła-Staśkowiak, J. (2019). A contrastive analysis of feminitives in Bulgarian, Polish and Russian. *Cognitive Studies | Études cognitives*, *2019*(19), Article 1922. https://doi.org/10.11649/cs.1922

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. W *Proceedings of the 19th Annual Conference of the European Association for Machine Translation (EAMT)* (s. 384). Baltic Journal of Modern Computing.

von Waldenfels, R. (2011). Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. W D. Majchráková & R. Garabík (Red.), *Natural language processing, multilinguality. Sixth international conference Modra, Slovakia, 20–21 October 2011: Proceedings* (ss. 156–162). Tribun EU.

Wawer, A. (2019). Sentiment analysis for Polish. *Poznań Studies in Contemporary Linguistics*, *55*(2), 445–468. http://doi.org/10.1515/psicl-2019-0016

Добровольский, Д., Кретов, А., & Шаров, С. (2005). Корпус параллельных текстов: Архитектура и возможности использования. W Д. Добровольский, А. Кретов, & С. Шаров, *Национальный корпус русского языка: 2003–2005* (ss. 263–296). Индрик.

Bibliography (Transliteration)

Blagoeva, D., Jaskot, M. P., & Sosnowski, W. (2019). A lexicographical approach to the contrastive analysis of Bulgarian and Polish phraseology. *Cognitive Studies | Études cognitives*, *2019*(19), Article 1923. https://doi.org/10.11649/cs.1923

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427. https://doi.org/10.1075/ijcl.17.3.05cer

Charciarek, A. (2018). Možnosti využiti korpusu InterCorp v česko-polske překladové lexikografii. *Časopis pro moderní filologii*, *100*(2), 206–222.

Charciarek, A. (2019a). Korpus równoległy InterCorp w leksykografii przekładowej – możliwości i ograniczenia. *Roczniki Humanistyczne*, *67*(7), 79–92. https://doi.org/10.18290/rh.2019.67.7-5

Charciarek, A. (2019b). Využití paralelního korpusu v translatologii (na základě česko-polského InterCorpu). *Bohemistyka*, *2019*(2), 194–216. https://doi.org/10.14746/bo.2019.2.5

CLARIN-PL. (n.d.). *Polska infrastruktura CLARIN*. http://clarin-pl.eu/

Dobrovolʹskiĭ, D., Kretov, A., & Sharov, S. (2005). Korpus parallelʹnykh tekstov: Arkhitektura i vozmozhnosti ispolʹzovaniia. In D. Dobrovolʹskiĭ, A. Kretov, & S. Sharov, *Natsionalʹnyĭ korpus russkogo iazyka: 2003–2005* (pp. 263–296). Indrik.

Duszkin, M., Roszko, D., & Roszko, R. (2021). New parallel corpora of Baltic and Slavic languages – Assumptions of corpus construction. In K. Ekštein, F. Pártl, & M. Konopík (Eds.), *Lecture Notes in Artificial Intelligence LNAI 12848: TSD 2021* (pp. 173–183). Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-83527-9_15

Dziob, A., & Piasecki, M. (2018). Dynamic verbs in the Wordnet of Polish. *Cognitive Studies | Études cognitives*, *2018*(18), Article 1718. https://doi.org/10.11649/cs.1718

Fellerer, J. (2020). *Urban multilingualism in East-Central Europe: The Polish dialect of late-Habsburg Lviv*. Rowman & Littlefield.

Garncarek, Ł., Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Turski, M., & Graliński, F. (2021). LAMBERT: Layout-aware language modeling for information extraction. In J. Lladós, D. Lopresti, & S. Uchida (Eds.), *Document Analysis and Recognition – ICDAR 2021* (pp. 532–547). Springer International Publishing. https://doi.org/10.1007/978-3-030-86549-8_34

Grabowski, Ł. (2012). *A corpus-driven study of translational and non-translational texts: The case of Nabokov's* Lolita. Wydawnictwo Uniwersytetu Opolskiego.

Grabowski, Ł. (2015). O frazeologii z perspektywy językoznawstwa korpusowego: Przegląd głównych nurtów badawczych ostatniego dwudziestolecia w Wielkiej Brytanii i USA. *Problemy Frazeologii Europejskiej*, *10*, 23–48.

Gruszczyńska, E., & Leńko-Szymańska, A. (Eds.). (2016). *Polskojęzyczne korpusy równoległe*. Instytut Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Hebal-Jezierska, M. (2013). Podstawowe zasady korzystania z korpusów przy badaniu języka. In W. Chlebda (Ed.), *Tropem korpusów: W poszukiwaniu optymalnych zbiorów tekstów* (pp. 17–30). Uniwersytet Opolski.

Hebal-Jezierska, M. (Ed.). (2014). *Praktyczny przewodnik po korpusach języków słowiańskich*. Wydział Polonistyki Uniwersytetu Warszawskiego.

Kaczmarska, E. (2019). *Metody ustalania ekwiwalentów czasowników wyrażających stany emocjonalne w przekładzie czesko-polskim na materiale z korpusu równoległego InterCorp*. Wydział Polonistyki Uniwersytetu Warszawskiego.

Kaźmierczak, I., Sarzyńska-Wawer, J., Wawer, A., & Chądzyńska, M. (2020). Describing a critical life event and its psychological consequences: The type of language used by patients suffering from depression and its relationship with personality development. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues.* https://doi.org/10.1007/s12144-020-00944-5

KonText – Corpus Query Interface. (n.d.). https://kontext.clarin-pl.eu/run.cgi/first_form

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2007). *Polsko-bułgarska gramatyka konfrontatywna*. Wydawnictwo Akademickie "Dialog".

Koseska-Toszewa, V., Korytkowska, M., & Roszko, R. (2009). Contrastive studies and semantic interlanguage. *Cognitive Studies | Études cognitives*, *2009*(9), 15–34.

Koseska, V., & Roszko, R. (2015). On semantic annotation in CLARIN-PL parallel corpora. *Cognitive Studies | Études cognitives*, *2015*(15), 211–236. https://doi.org/10.11649/cs.2015.016

Łaziński, M., & Kuratczyk, M. (2016). Korpus Polsko-Rosyjski Uniwersytetu Warszawskiego. In E. Gruszczyńska & A. Leńko-Szymańska (Eds.), *Polskojęzyczne korpusy równoległe* (pp. 83–95). Instytut Lingwistyki Stosowanej Uniwersytetu Warszawskiego.

Leech, G. (2002). The importance of reference corpora. In *Corpus lingüísticos: Presente y futuro* (pp. 1–11). Unibertsitate Zerbitzuetarako Euskal Ikastetxea.

Machálek, T. (2020). KonText: Advanced and flexible corpus query interface. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 7003–7008). European Language Resources Association.

Marcińczuk, M., Oleksy, M., Kocoń, J., Bernaś, T., & Wolski, M. (2015). Towards an event annotated corpus of Polish. *Cognitive Studies | Études cognitives*, *2015*(15), 253–267. https://doi.org/10.11649/cs.2015.018

Maziarz, M., & Rudnicka, E. (2020). Expanding WordNet with gloss and polysemy links for evocation strength recognition. *Cognitive Studies | Études cognitives*, *2020*(20), Article 2325. https://doi.org/10.11649/cs.2325

PELCRA Polish-Russian parallel corpus. (n.d.). http://pelcra.pl/new/polrus

Pęzik, P. (2020). Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*, *7*, 133–150. http://doi.org/10.31261/FL.2020.07.11

Piasecki, M., Walkowiak, T., Rudnicka, E., & Bond, F. (2018). Lexical platform – the first step towards user-centred integration of lexical resources. *Cognitive Studies | Études cognitives*, *2018*(18), Article 1811. https://doi.org/10.11649/cs.1811

Piotrowski, T., & Grabowski, Ł. (2013). Interpretacja danych frekwencyjnych z korpusów językowych: Opis pewnych problemów (na kilku przykładach z życia wziętych). In W. Chlebda (Ed.), *Na tropach korpusów: W poszukiwaniu optymalnych zbiorów tekstów* (pp. 59–71). Wydawnictwo Uniwersytetu Opolskiego.

Przepiórkowski, A., Bańko, M., Górski, R., & Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN.

Reynaert, M. (2006). Corpus-induced corpus clean-up. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC–2006, Trento* (pp. 87–92). European Language Resources Association (ELRA). https://aclanthology.org/L06-1000

Roszko, D. (2015). *Zagadnienia kwantyfikacyjne i modalne w litewskiej gwarze puńskiej (na tle literackich języków polskiego i litewskiego)*. Instytut Slawistyki Polskiej Akademii Nauk. https://hdl.handle.net/20.500.12528/31

Sosnowski, W. P., & Satoła-Staśkowiak, J. (2019). A contrastive analysis of feminitives in Bulgarian, Polish and Russian. *Cognitive Studies | Études cognitives*, *2019*(19), Article 1922. https://doi.org/10.11649/cs.1922

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19tb Annual Conference of the European Association for Machine Translation (EAMT)* (p. 384). Baltic Journal of Modern Computing.

von Waldenfels, R. (2011). Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In D. Majchráková & R. Garabík (Eds.), *Natural language processing, multilinguality. Sixth international conference Modra, Slovakia, 20–21 October 2011: Proceedings* (pp. 156–162). Tribun EU.

Wawer, A. (2019). Sentiment analysis for Polish. *Poznań Studies in Contemporary Linguistics*, *55*(2), 445–468. http://doi.org/10.1515/psicl-2019-0016

O nowych ręcznie zrównoleglonych i znakowanych dwujęzycznych korpusach równoległych oraz ich zastosowaniach

Streszczenie

W artykule autor opisuje obecnie powstające ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe CLARIN-PL-BIZ języków bałtyckich i słowiańskich. Omawia wyróżniające cechy tych korpusów, które sprawią, że zastosowania tych korpusów znacznie wykroczą poza typowe analizy korpusowe. Wśród zastosowań tych korpusów autor wymienia definiowanie modeli międzyjęzykowych na rzecz rozwoju przekładu maszynowego i rozwoju sztucznej inteligencji. Zwraca również uwagę na wysoki potencjał tych zasobów jako wzorcowej bazy treningowej do testowania narzędzi przetwarzania języka naturalnego.

Słowa kluczowe: ręcznie zrównoleglone i znakowane dwujęzyczne korpusy równoległe; język litewski; języki słowiańskie; narzędzia językowe; CLARIN-PL

On New Manually Aligned and Tagged Bilingual Parallel Corpora and Their Applications

Abstract

This article is devoted to the manually aligned and tagged bilingual parallel CLARIN-PL-BIZ corpora of the Baltic and Slavic languages which are currently being developed. The study discusses the essential features of these corpora that make their applications go far beyond typical corpus analysis. Applications of these corpora include the design of cross-language models for the development of machine translation and artificial intelligence. The article also draws attention to the high potential of these resources as a model training base for testing natural language processing tools.

Keywords: manually aligned and tagged bilingual parallel corpora; Lithuanian language; Slavic languages; language tools; CLARIN-PL

Dr hab. Roman Roszko, Associate Professor at the Institute of Slavic Studies, Polish Academy of Sciences, Warsaw; PhD – 1992, Institute of Slavic Studies, PAS; habilitation in linguistics – 2005, Institute of Slavic Studies, PAS; author or co-author of five books, over one hundred and twenty scholarly articles; editor or co-editor of over twenty multi-author monographs. Research interests: contrastive linguistics of Baltic and Slavic languages, theoretical semantic research with an interlanguage, corpus linguistics and natural language processing. Co-author of a Polish–Bulgarian contrastive grammar, author of two monographs on the semantic category of definiteness/indefiniteness and evidentiality in Polish and Lithuanian. Co-author of ten parallel corpora. Head or member of several research projects.

Bibliography (selected): Lexical exponents of hypothetical modality in Polish and Lithuanian, *Cognitive Studies | Études cognitives* 12, Warszawa (Warsaw) 2012, 15–25; (with D. Roszko), A net presentation of Lithuanian sentences containing verbal forms with the grammatical suffix *-dav-*, *Cognitive Studies | Études cognitives* 14, Warszawa (Warsaw) 2014, 173–182; (with L. Dimitrova, V. Koseska-Toszewa, & D. Roszko), Trilingual aligned corpus – Current state and new applications, *Cognitive Studies | Études cognitives* 14, Warszawa (Warsaw) 2014, 13–20; (with V. Koseska-Toszewa), On semantic annotation in CLARIN-PL parallel corpora, *Cognitive Studies | Études cognitives* 15, Warszawa (Warsaw) 2015, 211–236; (with M. Duszkin, & D. Roszko), New parallel corpora of Baltic and Slavic languages – Assumptions of corpus con-

struction, in: K. Ekštein, F. Pártl, & M. Konopík (eds.), *Text, speech, and dialogue*, Cham 2021, 172–183.

Correspondence: Roman Roszko, Institute of Slavic Studies, Polish Academy of Sciences, Warsaw, e-mail: roman.roszko@ispan.waw.pl

Support of the work: This work was supported by CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.

Competing interests: The author declares that he has no competing interests.

Acknowledgment: I would like to thank the reviewers for their valuable comments, which enabled me to improve this article for the benefit of readers.

Publication History: Received: 2021-05-20; Accepted: 2021-11-08; Published: 2021-12-21